AD-A082 178 CLEMSON UNIV S C DEPT OF MATHEMATICAL SCIENCES F/G 12/1
RIDGE ESTIMATION FOR THE LINEAR REGRESSION MODEL.(U)
FEB 77 K ALAM, J S HAWKES N00014-75-C-0451
UNCLASSIFIED N92 NL

1 OF 1
AD
A082178

END
DATE
FILMED
4-80
DTIC

6

RIDGE ESTIMATION FOR THE LINEAR REGRESSION MODEL.

10

KHURSHEED ALAM
JAMES S. HAWKES, III

15 NØØØ14-75-C-Ø451

12 21

11 18 Feb 77

A

14

REPORT N92, TR-240

FEBRUARY 18, 1977

9 TECHNICAL REPORT, #240

RESEARCH SUPPORTED IN PART BY
THE OFFICE OF NAVAL RESEARCH
TASK NR 042-271 CONTRACT N0014-75-C-0451

407123

RIDGE ESTIMATION FOR THE LINEAR REGRESSION MODEL.

James S. Hawkes and Khursheed Alam*
Clemson University

ABSTRACT

Consider the linear regression model $Y = X\Theta + \varepsilon$, where Θ is an unknown parameter vector to be estimated. A class of estimators, variously known as the ridge estimators, is given by $\hat{\Theta} = (X'X + KI)^{-1}X'Y$, where K is a constant or a function of Y. The ridge estimator is a suitable alternative to the least squares estimator when the design matrix $X'X$ is nearly singular. A number of papers has appeared in the statistical literature in the recent years, giving empirical evaluation of various ridge estimators. This paper gives a theoretical discussion of some properties of the ridge estimators.

Key words: Linear Model; Regression; Mean Squared error; Least Squares Estimator; Bayes Estimator; Ridge Estimator.

AMS Classification: 62J05

*The author's work was supported by the Office of Naval Research under Contract N00014-75-0451.

INTRODUCTION

Consider the linear regression model

$$Y = X\Theta + \varepsilon \tag{1.1}$$

where Y is n x 1 vector of observations, X is n x p design matrix of rank p, Θ is p x 1 vector of unknown parameters and ε is n x 1 vector of observational errors. Let the components of ε be uncorrelated and have zero mean and a common variance equal to σ^2, say. The usual estimator of θ is derived by the least squares method, that is, by minimizing $(Y-X\Theta)'(Y-X\Theta)$ with respect to Θ, and is given by

$$\hat{\Theta} = (X'X)^{-1}X'Y \tag{1.2}$$

where prime denotes the transpose of a matrix. Clearly, $\hat{\Theta}$ is an unbiased estimator of Θ. Let $\lambda_1,\ldots,\lambda_p$ denote the characteristic roots of X'X. The mean squared error (MSE) of $\hat{\Theta}$ is given by

$$\begin{aligned} \mathrm{MSE}\hat{\Theta} &= E(\hat{\Theta}-\Theta)'(\hat{\Theta}-\Theta) \\ &= \sigma^2\Sigma_{i=1}^{p} \frac{1}{\lambda_i}. \end{aligned} \tag{1.3}$$

In application of multiple linear regression, the design matrix X'X is often nearly singular. This is due to some interrelation between the explanatory variables. The relation is technically called multicollinearity. The least squares estimator of the regression coefficients tends to become "unstable" in the presence of multicollinearity. More precisely, the variance of the estimates of some of the regression coefficients becomes large. This is shown by (1.3). For this case Hoerl (1962) and Hoerl and Kennard

(1970, a), (1970, b) suggested a class of estimators known as ridge estimators as an alternative to the least squares estimator. The ridge estimator is given by

$$\tilde{\Theta} = (X'X + KI)^{-1}X'Y \qquad (1.4)$$

where I denotes an identity matrix and K is a positive number or a suitable function of Y. Clearly, $\tilde{\Theta}$ is a biased estimator of Θ. The new method of estimation is called ridge estimation.

Let P be an orthogonal matrix, diagonalizing X'X, that is

$$PX'X\,P' = D \qquad (1.5)$$

where D is a diagonal matrix with the ith diagonal element equal to λ_i. Let $\alpha = (\alpha_1,\ldots,\alpha_p)' = P\theta$. If K is a constant, the mean squared error of $\tilde{\Theta}$ is given by

$$\begin{aligned} MSE\tilde{\theta} &= E(\tilde{\theta}-\theta)'(\tilde{\theta}-\theta) \\ &= \sigma^2\Sigma_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+K)^2} + K^2\Sigma\frac{\alpha_i^2}{(\lambda_i+K)^2} . \end{aligned} \qquad (1.6)$$

Comparing (1.3) with (1.6) we observe that the effect of multicollinearity of the explanatory variables in the design matrix on the mean squared error is suitably reduced by the ridge estimation.

Applied statisticians have shown considerable interest in ridge estimation. Papers by Farebrother (1975), Hawkins (1975), Hemmerle (1975), Hoerl, Kennard and Baldwin (1975), McDonald (1975), McDonald and Galarneau (1975), Newhouse and Oman (1971) and Sidik (1975) may be cited for reference. Most of these papers deal with the empirical evaluation, based on simulation study, of various ridge estimators and its comparison with the least squares estimator and other biased estimators. Since a large number of variables is involved in the regression problem, the given empirical results do not give sufficient insight into the operating characteristics of the ridge estimators. This paper gives a theoretical discussion of an expository nature of the ridge estimation. Among other results it is shown that for a certain choice of K, depending on Y, the ridge estimator has uniformly smaller mean squared error than the least squares estimator, if a number of characteristic roots of the design matrix is sufficiently small.

A generalized ridge estimator is given by

$$\Theta_o = (X'X + K_o)^{-1}X'Y \qquad (1.7)$$

where K_o is a diagonal matrix. In this paper we consider only the ordinary ridge estimator, given by (1.4).

RIDGE ESTIMATION

The main results of the paper are given by the following theorems. First we give a derivation of the ridge estimator based on the least squares principle. A slightly different derivation based on the same principle was given by Hoerl and Kennard (1970 a). Let c be a positive number, and let

$$R(\Theta) = (y-X\Theta)'(Y-X\Theta).$$

Theorem 2.1. The value of Θ minimizing $R(\Theta)$, given $\Theta'\Theta \leq c$, is equal to $\tilde{\Theta}$, where K is chosen such that $\tilde{\Theta}'\tilde{\Theta} = c$.

Proof: From (1.4) and (1.5) we have

$$\tilde{\Theta}'\tilde{\Theta} = (PX'Y)' \ (D + KI)^{-2}(PX'Y). \qquad (2.1)$$

From (2.1) it is seen that $\tilde{\Theta}'\tilde{\Theta}$ is decreasing in K. Therefore, the value of K, given by $\tilde{\Theta}'\tilde{\Theta} = c$, is uniquely determined.

We have

$$\begin{aligned} R(\tilde{\Theta}) &= (Y-X\tilde{\Theta})'(Y-X\tilde{\Theta}) \\ &= (Y-X\hat{\Theta})'(Y-X\hat{\Theta}) + (X'Y)'[(X'X + KI)^{-1}-(X'X)^{-1}] \\ &\qquad X'X[(X'X + KI)^{-1}-(X'X)^{-1}]X'Y \\ &= (Y-X\hat{\Theta})'(Y-X\hat{\Theta}) + (PX'Y)'D^*(PX'Y) \qquad (2.2) \end{aligned}$$

where D* is a pxp diagonal matrix whose ith diagonal element is equal to

$$\frac{K^2}{\lambda_i(K+\lambda_i)^2}$$

It is seen from (2.2) that $R(\tilde{\Theta})$ is increasing in K.

Now, consider the problem of minimizing $R(\Theta)$ with respect to Θ under the constraint $\Theta'\Theta = c$. By the Lagrangian method the minimizing value of Θ is given by

$$\lambda\Theta-X'(Y-X\Theta) = o$$

or

$$\Theta = (X'X + \lambda I)^{-1}X'Y$$

where λ is determined such that $\Theta'\Theta = c$. Thus $R(\Theta)$ is minimized for $\Theta = \tilde{\Theta}$, where K is determined such that $\tilde{\Theta}'\tilde{\Theta} = c$.

We have shown above that $R(\tilde{\Theta})$ is increasing in K and that $\tilde{\Theta}'\tilde{\Theta}$ is decreasing in K. It follows that $\tilde{\Theta}$ which is the minimizing value of $R(\Theta)$, given $\Theta'\Theta = c$, is also the minimizing value of $R(\Theta)$, given $\Theta'\Theta \leq c$, where K is determined from $\tilde{\Theta}'\tilde{\Theta} = c$. □

Remark 1. The above theorem gives an interesting comparison between the derivation of the least squares estimator and the ridge estimator. The ridge estimator is derived by minimizing $R(\Theta)$ under a certain constraint on the value of $\Theta'\Theta$, whereas the least squares estimator is derived by minimizing $R(\Theta)$ without that constraint.

The next theorem gives another derivation of the ridge estimator from a Bayesian approach, assuming that the prior distribution of Θ and the conditional distribution of Y given Θ, are both normal. The proof of the theorem is trivial. This result is also noted by Lindley and Smith (1972). The result implies that the ridge estimator for a constant value of K is a Bayes estimator and admissible under squared error loss. The notation $Y \overset{d}{\sim} N(\mu,\Sigma)$ means that Y has a (multivariate) normal distribution with mean vector μ and covariance matrix Σ.

Theorem 2.2. If $Y \overset{d}{\sim} N(X\Theta,\sigma^2 I)$ conditionally given Θ, and a' priori $\Theta \overset{d}{\sim} N(0,\tau^2 I)$ then the posterior mean of Θ given Y, is equal to $\tilde{\Theta}$ for $K = \sigma^2/\tau^2$.

It is natural to compare the mean squared error of the ridge estimator and the least squares estimator. First we consider the case when K is a constant. It is clear from

(1.3) and (1.6) that for any $K > o$

$$\text{MSE}\tilde{\Theta} > \text{MSE}\hat{\Theta}$$

for sufficiently large value of $\Theta'\Theta$. On the other hand, if it is known a' priori that $\Theta'\Theta \leq c$ for some positive number c, a valid condition in many practical situations, then from (1.6) we get

$$\text{MSE}\tilde{\Theta} \leq \sigma^2 \Sigma_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + K)^2} + cK^2 \Sigma_{i=1}^{p} \frac{1}{(\lambda_i + K)}. \qquad (2.3)$$

Theorems 2.3 and 2.4 below, give values of K obtained from (2.3), for which the ridge estimator has smaller mean squared error than the least squares estimator.

<u>Theorem 2.3</u>. If $\Theta'\Theta \leq c$ then $\text{MSE}\tilde{\Theta} < \text{MSE}\hat{\Theta}$ for $o < K \leq \frac{2\sigma^2}{c}$.

Proof: From (2.3) we have for $o < K \leq \frac{2\sigma^2}{c}$

$$\text{MSE }\tilde{\Theta} \leq \sigma^2 \Sigma_{i=1}^{p} \left(\frac{\lambda_i}{(\lambda_i+K)^2} + \frac{2K}{(\lambda_i+K)^2}\right)$$

$$= \sigma^2 \; \Sigma_{i=1}^{p} \; \frac{\lambda_i + 2K}{(\lambda_i+K)^2} \,.$$

$$< \sigma^2 \; \Sigma_{i=1}^{p} \; \frac{1}{\lambda_i}$$

$$= \text{MSE } \hat{\Theta}. \; \square$$

<u>Theorem 2.4</u>. If $\Theta'\Theta < \frac{\sigma^2}{p} \Sigma_{i=1}^{p} \frac{1}{\lambda_i}$ then $\text{MSE}\tilde{\Theta} < \text{MSE}\hat{\Theta}$ for $K > o$.

Proof: Let $D(K)$ denote the quantity on the right hand side of (2.3). Differentiating $D(K)$ with respect to K we get

$$\partial D(K)/\partial K = \Sigma_{i=1}^{p} \frac{2\lambda_i (cK - \sigma^2)}{(\lambda_i + K)^3} \qquad (2.4)$$

The right hand side of (2.4) is equal to zero for $K = \sigma^2/c$ and is $<(>)o$ for $K < (>)\sigma^2/c$. Hence, $D(K)$ is first decreasing then increasing as K varies from o to ∞. Now, $D(\infty) = pc$ and

$$D(o) = \sigma^2 \Sigma_{i=1}^{p} \frac{1}{\lambda_i}$$

$$= \mathrm{MSE}\hat{\Theta}.$$

Therefore

$$\begin{aligned} \mathrm{MSE}\tilde{\Theta} &\leq D(K) \\ &\leq \max(pc, \mathrm{MSE}\hat{\Theta}) \\ &= \mathrm{MSE}\hat{\Theta} \end{aligned}$$

for $c \leq \frac{\sigma^2}{p} \Sigma_{i=1}^{p} \frac{1}{\lambda_i}$. □

An expression for K minimizing $\mathrm{MSE}\tilde{\Theta}$, given by (1.6), is not obtainable in a closed form. But for a given value of $\Theta'\Theta$, it is seen from (1.6) that $\mathrm{MSE}\tilde{\Theta}$ is minimized (maximized) by setting $\alpha_i^2 = \Theta'\Theta$ for the value of i corresponding to the largest (smallest) characteristic root of the matrix X'X. That is, $\mathrm{MSE}\tilde{\Theta}$ is minimized (maximized) for the value of θ proportional to the characteristic vector of X'X corresponding to the largest (smallest) characteristic root of the matrix. This result was also noted by Newhouse and Oman (1971). Let $\lambda_* = \min(\lambda_1, \ldots, \lambda_p)$ and

$$Q_c(K) = \sigma^2 \Sigma_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+K)2} + \frac{cK^2}{(\lambda_*+K)2} .$$

The following theorem follows from (1.6).

Theorem 2.5. A value of K minimizing $Q_c(K)$ is minimax for $\mathrm{MSE}\tilde{\Theta}$, given $\Theta'\Theta \leq c$.

Now we consider the case when K depends on Y. In this case, the main question is what is a suitable choice of K as a function of Y? Some of the authors cited above, have considered various choices and have compared the corresponding estimators with other estimators. Their comparison is mainly based on simulation study which leaves many questions unanswered.

In particular, it is not known whether the ridge estimator for any of those choices of K has smaller mean squared error than the least squares estimator for all values of Θ. We consider the choice of K, given by

$$K^* = \nu\hat{\sigma}^2/(\hat{\theta}'\hat{\theta}) \tag{2.5}$$

where ν is a positive number and

$$\hat{\sigma}^2 = \frac{Y'(I-X(X'X)^{-1}X')Y}{n-p} .$$

The given choice of K is suggested by Theorem 2.2, since $\hat{\sigma}^2$ is an unbiased estimate of σ^2 and $(\hat{\theta}'\hat{\theta})/p$ is an estimate of τ^2. Let

$$\Theta^* = (X'X+K^*I)^{-1}X'Y \tag{2.6}$$

denote the corresponding ridge estimator. Assuming that $Y \overset{d}{\sim} N(X\theta,\sigma^2 I)$, we shall compute MSE$\Theta^*$ and compare it with MSE$\hat{\Theta}$. The normality assumption will be made tacitly throughout the following discussion. Under the normality assumption, $\hat{\sigma}^2$ and $\hat{\theta}$ are independently distributed.

It is in order at this point to consider briefly the question of the inadmissibility of the least squares estimator with respect to a certain class of biased estimators. Since Stein (1955) showed that the mean of a p-variate normal distribution is inadmissible for $p \geq 3$, a large number of papers has been written on the subject. Alam (1973, 1975), Baranchik (1973), Berger (1976), Bhattacharya (1966), Bock (1975) and Sclove (1968), to name only a few, have considered certain class of estimators of the mean of the distribution, which dominate the least squares estimator. From Theorem 5 of Bock (1975) it follows that an estimator of the form

$$\delta = f\left(\frac{Y'X(X'X)^{-1}X'Y}{S}\right)\hat{\theta} \tag{2.7}$$

has a smaller MSE than $\hat{\Theta}$ for all values of Θ, where S is a random variable independent of X'Y, such that (S/σ^2) has a chi-square distribution with m degrees of freedom, $f: [o,\infty]\to[o,1]$, $y(1-f(y))$ is nondecreasing in y, $o\leq y(1-f(y))\leq (2\alpha-4)/(m+2)$ and

$$2<\alpha = (\Sigma_{i=1}^{p} \frac{1}{\lambda_i})\lambda_* . \tag{2.8}$$

The random variable S is given, for example, by

$$S = Y'(I-X(X'X)^{-1}X')Y$$

where m = n-p. By Theorem 6 of Bock, the inequality (2.8) is also necessary for an estimator of the form (2.7) to have smaller MSE than the least squares estimator. Clearly, the inequality holds for $p \geq 3$ if X'X is a constant multiple of the identity matrix. In this case and only in this case the ridge estimator is a multiple of $\hat{\Theta}$.

Now we compute the mean squared error of the ridge estimator Θ^*, given by (2.6). Let $\chi^2_{m,\gamma}$ denote a non-central chi-square random variable with m-degrees of freedom and non-centrality parameter γ. Let ϕ be an integrable function, and let $T \overset{d}{\sim} N(\xi,1)$. It is easily shown that

$$ET\phi(T^2) = \xi E\phi(\chi^2_{3,\xi^2}) \tag{2.9}$$

$$ET^2\phi(T^2) = E\phi(\chi^2_{3,\xi^2}) + \xi^2 E\phi(\chi^2_{5,\xi^2}) . \tag{2.10}$$

Let $Z = (Z_1,\ldots,Z_p)' = PX'Y$, where P is given by (1.5). We have that $Z \overset{d}{\sim} N(D\alpha,\sigma^2 D)$, $\hat{\theta}'\hat{\theta} = \Sigma_{i=1}^{p} Z_i^2/\lambda_i^2$ and

$$\begin{aligned} MSE\theta^* &= E(\theta^*-\theta)'(\theta^*-\theta) \\ &= E((D+K^*I)^{-1}Z-\alpha)'((D+K^*I)^{-1}Z-\alpha) \\ &= E\Sigma_{i=1}^{p}\left(\frac{Z_i}{\lambda_i+\nu\hat{\sigma}^2/(\Sigma_{j=1}^{p} Z_j^2/\lambda_j^2)} - \alpha_i\right)^2 \\ &= E\Sigma_{i=1}^{p}\left(\frac{\sqrt{\lambda_i}\,\sigma U_i}{\lambda_i+(\frac{\nu\hat{\sigma}^2}{\sigma^2})/(\Sigma_{j=1}^{p} U_j^2/\lambda_j)} - \alpha_i\right)^2 \end{aligned} \tag{2.11}$$

where $U_i = Z_i/(\sqrt{\lambda}_i\sigma) \underset{\sim}{d} N(\sqrt{\lambda}_i\alpha_i/\sigma, 1)$, so that $U_i^2 \underset{\sim}{d} \chi^2_{1,\lambda_i\alpha_i^2/\sigma^2}$. Let $V_i^2 \underset{\sim}{d} \chi^2_2, W_i^2 \underset{\sim}{d} \chi^2_2 (i=1,\ldots, p)$ independent of the U_i's and among themselves. Using (2.9) and (2.10) in (2.11) we get

$$\text{MSE}\,\theta^* = E\Sigma_{i=1}^p \frac{\lambda_i\ \sigma^2}{(\lambda_i + (\frac{\nu\hat{\sigma}^2}{\sigma 2})/(\Sigma_{j=1}^p \frac{U_j^2}{\lambda_j} + \frac{V_i^2}{\lambda_i}))^2}$$

$$+E\Sigma_{i=1}^p\ \alpha_i^2\ [\ \frac{\lambda_i^2}{(\lambda_i + (\frac{\nu\hat{\sigma}^2}{\sigma 2})/(\Sigma_{j=1}^p \frac{U_j^2}{\lambda_j} + \frac{V_i^2 + W_i^2}{\lambda_j}))^2}$$

$$- \frac{2\lambda_i}{\lambda_i + (\frac{\nu\ \hat{\sigma}^2}{\sigma 2})/(\Sigma_{j=1}^p \frac{U_j^2}{\lambda_j} + \frac{V_i^2}{\lambda_i})} + 1\ ]. \qquad (2.12)$$

It is not possible to simplify further the expression for MSEθ^*, except for some special cases. Therefore, we shall consider only those cases.

First suppose that the λ_i's are all equal to λ, say. From (2.12) we obtain after simplification

$$\text{MSE}\,\theta^* = \frac{\sigma^2}{\lambda}\ E[p(\frac{F}{F + \nu/(p+2)})^2 + \tau((\frac{F^*}{F^*+\nu/(p+4)})^2 - \frac{2F}{F+\nu/(p+2)} + 1)] \qquad (2.13)$$

where $F \underset{\sim}{d} (n-p)\chi^2_{p+2,\tau}/(p+2)\chi^2_{n-p}$, $F^* \underset{\sim}{d} (n-p)\chi^2_{p+4,\tau}/(p+4)\chi^2_{n-p}$ and $\tau = \lambda\alpha'\alpha/\sigma^2$.

The above expression for MSEθ^* is computable by integration from the density function of the non-central F-distribution. For $\tau = o$ we have

$$\text{MSE}\,\theta^* = \frac{p\sigma^2}{\lambda}\ E(\frac{\chi^2_{p+2}}{\chi^2_{p+2} + \nu\chi^2_{n-p}/(n-p)})^2$$

$$< \frac{p\sigma^2}{\lambda} = \text{MSE}\,\hat{\vartheta}.$$

For large values of τ we have

$$\text{MSE}\,\theta^* \approx \frac{\sigma^2}{\lambda}\, E[p(1-\frac{2\nu\chi^2_{n-p}}{(n-p,\chi^2_{p+2,\tau}}) + \tau(\frac{\nu\chi^2_{n-p}}{(n-p)\chi^2_{p+2,\tau}})^2]$$

$$= \frac{\sigma^2}{\lambda}\,[p-2p\nu E(\chi^2_{p+2,\tau})^{-1}+\frac{\tau\nu^2(n-p+2)}{n-p}\, E(\chi^2_{p+2,\tau})^{-2}]$$

$$= \frac{\sigma^2}{\lambda}\,[p-\frac{p\nu\,\Gamma(p/2)}{\Gamma(\frac{p}{2}+1)}\Phi(\frac{p}{2}, \frac{p}{2}+1; \frac{\tau}{2})e^{-\tau/2}$$

$$+\frac{\tau\nu^2(n-p+2)\Gamma(\frac{p}{2}-1)}{4(n-p)\,\Gamma(\frac{p}{2}+1)}\,\Phi\,(\frac{p}{2}-1,\frac{p}{2}+1; \frac{\tau}{2})e^{-\tau/2}]$$

$$\approx \frac{\sigma^2}{\lambda}[p-\frac{\nu}{\tau}(2p-\frac{\nu(n-p+2)}{n-p})]$$

$$= \text{MSE}\,\hat{\theta} - \frac{\nu\sigma^2}{\lambda\tau}\,(2p-\frac{\nu(n-p+2)}{n-p})$$

$$< \text{MSE}\,\hat{\theta} \text{ for } \nu<2p(n-p)/(n-p+2) \qquad (2.14)$$

where

$$\Phi(a,b;x) = 1 + \frac{a}{b}\,x + \frac{a(a+1)}{b(b+1)}\,\frac{x^2}{2!} + \ldots$$

denotes the confluent hypergeometric function. Since θ^* is of the form (2.7), an application of Bock's result shows that MSE $\theta^* \leq$ MSE $\hat{\theta}$ for all values of Θ if

$$\nu\leq 2(p-2)(n-p)/(n-p+2).$$

Next suppose that $\lambda_*\to 0$ and the remaining (p-1) characterist roots are bounded away from zero. Let $\lambda_j = \lambda_*$, $G \overset{d}{\sim}(n-p)\chi^2_3/3\chi^2_{n-p}$ and $G^* \overset{d}{\sim}(n-p)\chi^2_5/5\chi^2_{n-p}$. If $\lambda^*\alpha'\alpha\to 0$ then from (2.12) the value of MSE θ^* is approximated by

$$\text{MSE}\,\theta^* \approx \frac{\sigma^2}{\lambda_*}\, E(\frac{G}{G+\nu/3})^2+\Sigma_{i\neq j}\,\frac{\sigma^2}{\lambda_i}$$

$$+\alpha_j^2[E(\frac{G^*}{G^*+\nu/5})^2-2E(\frac{G}{G+\nu/3}) + 1]. \qquad (2.15)$$

Hence

$$\lambda_*(\mathrm{MSE}\hat{\theta}-\mathrm{MSE}\theta^*)\to\sigma^2(1-E(\frac{G}{G+\nu/3})^2). \qquad (2.16)$$

We have shown the following result.

Theorem 2.6. If the λ_i's are equal to λ, say, then

$$\lim_{\theta'\theta\to\infty}(\mathrm{MSE}\hat{\theta}-\mathrm{MSE}\theta^*)\theta'\theta = \frac{\nu\sigma^4}{\lambda^2}(2p-\frac{\nu(n-p+2)}{n-p})$$

and $\mathrm{MSE}\theta^* \leq \mathrm{MSE}\hat{\theta}$ for $\nu \leq 2(p-2)(n-p)/(n-p+2)$ for all values of θ. If $\lambda_*\to 0$ but the other p-1 values of the λ_i's are bounded away from zero, and $\lambda_*\alpha'\alpha\to 0$ then

$$\lim_{\lambda_*\to 0}\lambda_*(\mathrm{MSE}\hat{\theta}-\mathrm{MSE}\theta^*) = \sigma^2(1-E(\frac{G}{G+\nu/3})^2).$$

Suppose that $\lambda_i=\lambda_*$ for r values of i and the other values of the λ_i's are bounded away from zero. The following theorem shows that $\mathrm{MSE}\theta^*< \mathrm{MSE}\hat{\theta}$ for sufficiently small values of λ_*. The proof is based on certain results given in the Appendix.

Theorem 2.7. If $o<\nu<1$ and $r>4+\nu(n-p+2)/(n-p)$ then for λ_* sufficiently small $\mathrm{MSE}\theta^*<\mathrm{MSE}\hat{\theta}$ for all values of θ.

$\mathrm{MSE}\tilde{\theta}$ is bounded for any value of θ as $\lambda_*\to 0$ but $\mathrm{MSE}\theta^*\to\infty$. An alternative estimator for which the MSE is bounded, is given by

$$\theta^{**} = (X'X+K^{**}I)^{-1}X'Y$$

Where $K^{**} = \hat{\sigma}^2/\tilde{\theta}'\tilde{\theta}$. The mean squared error of θ^{**} is given by

$$\mathrm{MSE}\theta^{**} = E\Sigma_{i=1}^{p}\frac{\lambda_i\sigma 2}{(\lambda_i+(\hat{\sigma}^2/\sigma^2)/\Sigma_{j=1}^{p}\frac{\lambda_j U_j^2}{(\lambda_j+K)^2}+\frac{\lambda_i V_i^2}{(\lambda_i+K)^2}))^2}$$

$$+ E\Sigma_{i=1}^{p}\alpha_i^2[\frac{\lambda_i^2}{(\lambda_i+(\hat{\sigma}^2/\sigma^2)/(\Sigma_{j=1}^{p}\frac{\lambda_j U_j^2}{(\lambda_j+K)^2}+\frac{\lambda_i(V_i^2+W_i^2)}{(\lambda_i+K)^2}))^2}$$

$$-\frac{2\lambda_i}{\lambda_i+(\hat{\sigma}^2/\sigma^2)/(\Sigma_{j=1}^{2}\frac{\lambda_j U_j^2}{(\lambda_j+K)^2}+\frac{\lambda_i V_i^2}{(\lambda_i+K)^2})}+1]. \qquad (2.17)$$

corresponding to (2.12) for $\mathrm{MSE}\theta^*$.

APPENDIX

Proof of Theorem 2.7. Let $\lambda_1=\lambda_2=\ldots=\lambda_r=\lambda_*$ and let λ_i be bounded away from zero for $i>r$. We consider the limiting value of MSEθ^* as $\lambda_*\to 0$. Let B_i denote the quantity inside the square bracket in (2.12). Clearly, $0<B_i<2$. Also, $B_i\to 0$ as $\lambda_*\to 0$ for $i>r$. Therefore, the second summation in (2.12), given $\alpha'\alpha$, is maximized in the limiting case as $\lambda_*\to 0$ by putting $\alpha_i^2=0$ for $i=r+1,\ldots,p$. Similarly, the first summation in (2.12) is minimized in the limiting case as $\lambda_*\to 0$ by the same substitution. Therefore, we let $\alpha_i^2=0$ for $i=r+1,\ldots,p$.

From (2.12) the value of MSEθ^* as $\lambda_*\to 0$ is approximated by

$$\text{MSE}\,\theta^* \approx \frac{r\sigma^2}{\lambda_*} E(1+(\frac{\nu\hat{\sigma}^2}{\sigma2})/\chi^2_{r+2,\lambda_*\alpha'\alpha/\sigma^2})^{-2}+\sigma^2\Sigma_{i=r+1}^{p}\frac{1}{\lambda_i}$$

$$+\ \alpha'\alpha E[(1+(\frac{\nu\hat{\sigma}^2}{\sigma2})/\chi^2_{r+4,\lambda_*\alpha'\alpha/\sigma^2})^{-2}2(1+(\frac{\nu\hat{\sigma}^2}{\sigma2})/\chi^2_{r+2,\lambda_*\alpha'\alpha/\sigma^2})^{-1}] \tag{3.1}$$

where the χ^2random variables are distributed independent of $\hat{\sigma}^2$.

Let $\delta=\lambda_*\alpha'\alpha/\sigma^2$, $V \overset{d}{\sim} \chi^2_2$ and let Q denote the quantity inside the square bracket in (3.1). We have

$$E\,Q = E[(1-\frac{\nu\hat{\sigma}^2/\sigma^2}{V+\chi^2_{r+2,\delta}+\nu\hat{\sigma}^2/\sigma^2})^2-2(1-\frac{\nu\hat{\sigma}^2/\sigma^2}{\chi^2_{r+2,\delta}+\nu\hat{\sigma}^2/\sigma^2})+1]$$

$$= E[\frac{2(\nu\hat{\sigma}^2/\sigma^2)V}{(V+\chi^2_{r+2,\delta+\nu\hat{\sigma}^2/\sigma^2})(\chi^2_{r+2,\delta}+\nu\hat{\sigma}^2/\sigma^2)}+\frac{(\nu\hat{\sigma}^2/\sigma^2)^2}{(V+\chi^2_{r+2,\delta}+\nu\hat{\sigma}^2/\sigma^2)^2}]$$

$$\leq E[\frac{2(\nu\hat{\sigma}^2/\sigma^2)E(V)}{(\chi^2_{r+2,\delta}+\nu\hat{\sigma}^2/\sigma^2)^2}+\frac{(\nu\hat{\sigma}^2/\sigma^2)^2}{(\chi^2_{r+2,\delta}+\nu\hat{\sigma}^2/\sigma^2)^2}]$$

$$= E(\nu\hat{\sigma}^2/\sigma^2)(4+\nu\hat{\sigma}^2/\sigma^2)(\chi^2_{r+2,\delta}+\nu\hat{\sigma}^2/\sigma^2)^{-2}$$

$$= \nu E(4+\nu\chi^2_{n-p+2}/(n-p))(\chi^2_{r+2,\delta}+\nu\chi^2_{n-p+2}/(n-p))^{-2}$$

$$\leq \nu E(4+\nu\chi^2_{n-p+2}/(n-p))E(\chi^2_{r+2,\delta}+\nu\chi^2_{n-p+2}/(n-p))^{-2}$$

$$= \nu(4+\nu\frac{n-p+2}{n-p})E(\chi^2_{r+2,\delta}+\nu\chi^2_{n-p+2}/(n-p))^{-2}. \qquad (3.2)$$

For the first term in (3.1) we have

$$E(1+\frac{\nu\hat{\sigma}^2}{\sigma^2}/\chi^2_{r+2,\delta})^{-2} \leq 1-E(\nu\hat{\sigma}^2/\sigma^2)(\chi^2_{r+2,\delta}+\nu\hat{\sigma}^2/\sigma^2)^{-1}$$

$$= 1-\nu E(\chi^2_{r+2,\delta}+\nu\chi^2_{n-p+2}/(n-p))^{-1} \qquad (3.3)$$

Combining (3.2) and (3.3), we get an upper bound on the right hand side of (3.1), given by

$$MSE\hat{\theta}+\frac{\nu\sigma^2}{\lambda_*}[\delta(4+\nu\frac{n-p+2}{n-p})E(\chi^2_{r+2,\delta}+\nu\chi^2_{n-p+2}/(n-p))^{-2}$$

$$- rE(\chi^2_{r+2,\delta}+\nu\chi^2_{n-p+2}/(n-p))^{-1}]. \qquad (3.4)$$

The fifth line in (3.2) and the second line in (3.3) is obtained from the relation $E\,\chi^2_m\,\phi(\chi^2_m) = m\phi(\chi^2_{m+2})$ for any integrable function ϕ.

Let R denote the quantity inside the square bracket in (3.4). Clearly, $R < o$ for sufficiently small values of δ. On the other hand, if $\delta\to\infty$ then

$$R \approx (4+\frac{\nu(n-p+2)}{n-p}-r)/\delta$$

$$< o \text{ for } r>4+\frac{\nu(n-p-2)}{n-p}.$$

If $\nu=o$ then

$$R = 4\,\delta\,E\,(\chi^2_{r+2,\delta})^{-2}-r\,E(\chi^2_{r+2,\delta})^{-1}$$

$$= [\frac{\delta\Gamma(\frac{r}{2}-1)}{\Gamma(\frac{r}{2}+1)}\Phi(\frac{r}{2}-1,\frac{r}{2}+1;\frac{\delta}{2})-\Phi(\frac{r}{2},\frac{r}{2}+1;\frac{\delta}{2})]e^{-\delta/2}$$

$$= [\frac{4\delta}{r(r-2)}\Phi(\frac{r}{2}-1,\frac{r}{2}+1;\frac{\delta}{2})-\Phi(\frac{r}{2},\frac{r}{2}+1;\frac{\delta}{2})]e^{-\delta/2}. \qquad (3.5)$$

Using the recurrence relation

$$b\Phi(a,b;x)-b\Phi(a-1,b;x) = x\Phi(a,b+1;x)$$

and the integral representation formula

$$\frac{\Gamma(b-a)\Gamma(a)}{\Gamma(b)}\ \Phi(a,b;x)=\int_o^1 e^{xt}t^{a-1}(1-t)^{b-a-1}dt,\ b>a>o$$

it can be shown that the value of R, given by (3.5) is negative for $r>4$.

If $\nu=n-p$ then

$$\begin{aligned} R &= \delta(n-p+6)E(\chi^2_{n-p+r+4,\delta})^{-2}-r\ E(\chi^2_{n-p+r+4,\delta})^{-1} \\ &= [\frac{\delta(n-p+6)\Gamma((n-p+r)/2)}{4\Gamma((n-p+r+4)/2)}\ \Phi(\frac{n-p+r}{2},\ \frac{n-p+r+4}{2};\ \frac{\delta}{2}) \\ &\quad - \frac{r\Gamma((n-p+r+2)/2)}{2\Gamma((n-p+r+4)/2)}\ \Phi\ (\frac{n-p+r+2}{2},\ \frac{n-p+r+4}{2};\ \frac{\delta}{2})]e^{-\delta/2}. \end{aligned} \tag{3.6}$$

As for (3.5) it can be shown that (3.6) is negative for $r>(n-p+6)$.

The above result suggests that the value of R is negative for all δ, and therefore $MSE\theta^*<MSE\hat{\theta}$ if $r>4+\frac{\nu(n-p+2)}{n-p}$. This result is connistent with the numerical values of R which have been computed for $\delta=1(1)5,10,15$ $\nu=2(.2)1.0$, $n-p = 5(5)25$ and $r=5\ (1)10$.

References

[1] Alam, K. (1973). A family of admissible minimax estimators of the mean of a multivariate normal distribution. Ann. Statist. (1) 517-25.

[2] Alam, K. (1975). Minimax and admissible minimax estimators of the mean of a multivariate normal distribution for unknown covariance matrix. J. Multivariate Anal. (5) 83-95.

[3] Baranchik, A. J. (1973). Inadmissibility of miximum likelihood estimators in some multiple regression problems with three or more independent variables. Ann. Statist. (1) 312-21.

[4] Berger, J. (1976). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. Ann. Statist. (4) 223-6.

[5] Bhattacharya, P. K. (1966). Estimating the mean of a multivariate normal population with general quadratic loss function. Ann. Math. Statist. (37) 1819-24.

[6] Bock, M. E. (1975). Minimax estimators of the mean of a multivariate normal distribution. Ann. Statist. (3) 209-18.

[7] Farebrother, R. W. (1975). The minimum mean square error linear estimator and ridge regression. Technometrics (17) 127-8.

[8] Hawkins, D. M. (1975). Relations between ridge regression and eigenanalysis of the augmented correlation matrix. Technometrics (17) 477-80.

[9] Hemmerle, W. J. (1975). An explicit solution for generalized ridge regression. Technometrics (1975) 309-14.

[10] Hoerl, A. E. (1962). Applications of ridge analysis to regression problems. Chemical Engineering Progress (58) 54-59.

[11] Hoerl, A. E. and Kennard, R. W. (1970 a). Biased estimation for non-orthogonal problems. Technometrics (12) 55-67.

[12] Hoerl, A. E. and Kennard, R. W. (1970 b). Ridge regression: applications to non-orthogonal problems. Technometrics (12) 69-82.

[13] Hoerl, A. E., Kennard, R. W. and Baldwin, K. F. (1975). Ridge regression: some simulations. Communications In Statistics (4) 105-23.

[14] Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *Jour. Royal Stat. Soc. Series B*. (34) 1-18.

[15] McDonald, G. C. and Galarneau, D. I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *Jour. Amer. Stat. Assoc*. (70) 407-16.

[16] Newhouse, J. P. and Oman, S. D. (1971). An evaluation of ridge estimators. *Rand Technical Report* R-716-PR.

[17] Sclove, S. L. (1968). Improved estimators for coefficients in linear regression. *Jour. Amer. Stat. Assoc*. (63) 597-606.

[18] SideK, S. M. (1975). Comparison of some biased estimation methods (including ordinary subset regression) in the linear model. *NASA Technical Report* TND-7932.

[19] Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Prob*. (1). 197-206.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
N- 92 ✓

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Ridge estimation for the linear regression model

5. TYPE OF REPORT & PERIOD COVERED
TR*240 ✓

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
James S. Hawkes
Khursheed Alam

8. CONTRACT OR GRANT NUMBER(s)
N00014-75-C-0451

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Clemson University
Dept. of Mathematical Sciences ✓
Clemson, South Carolina 29631

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
NR 042-271

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
Code 436
Arlington, Va. 22217

12. REPORT DATE
2/77

13. NUMBER OF PAGES
18

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Linear Model; Regression; Mean squared error; Least Squares; Bayes Estimator; Ridge Estimator

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A class of estimators, variously known as ridge estimators, is considered for the linear regression model $Y=X\beta+\epsilon$, where β is an unknown parameter vector to be estimated. Some properties of the ridge estimators are given. It is shown that certain ridge estimators have uniformly smaller mean squared error than the least squares estimator.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)